A Feasibility Study of Challenges and Opportunities in Computational Biology: A Malaysian Perspective

R.J.A. Richard and N. Sriraam
Center for Multimedia Computing, Faculty of Information Technology
Multimedia University, Cyberjaya, Malaysia

Abstract: The term computational biology refers to the knowledge derived from a computer analysis of biological data that includes identification of genes in DNA sequence of different organisms, prediction of structural and functional mechanism of proteins, feature extraction and classification of genomics and proteomics. Computational biology is a rapidly developing branch of science and is highly interdisciplinary, using techniques and Concepts from informatics, mathematics, chemistry, physics, statistics and biochemistry. This field has risen in parallel with the developments of automated high throughput methods of biochemistry and biological discovery that yield a variety of forms of experimental data, such as DNA& RNA sequences, gene expressions patterns and chemical structures. The field's rapid growth is spurred by the vast potential for new understanding that can lead to new technological treatments, new agro-crops cultivation and new pharmaceutical drug discovery. In the recent years, most Bioengineering disciplines are started adopting the information technology oriented curriculum due to its high performance computing, data interoperability, web-based platform compatibility and secured a suitable job opportunity. This study discusses the challenges to set up an interdisciplinary oriented curriculum by merging life sciences and information technology at a university level. It also provides the career opportunities for different life science disciplines like drug development, microbial genome applications, biotechnology, forensic and analysis of microbes.

Key words: Computational biology, genomics, proteomics, interdisciplinary oriented curriculum

INTRODUCTION

The term Computational biology refers to the knowledge derived from a computer analysis of biological data that includes identification of genes in DNA sequence of different organisms, prediction of structural and functional mechanism of proteins, feature extraction and classification of genomics and proteomics. Computational biology is a rapidly developing branch of science and is highly interdisciplinary, using techniques and concepts from informatics, mathematics, chemistry, physics, statistics and biochemistry[1-5]. Computational Biology today consists of a collection of methods and computational approaches and tools that have been developed in very disparate areas to analyze genomic data. This technology has been evolved to provide functional genomics support in terms of data management, the decision process, database integration and access to many of the large databases being produced with today's technological innovations[2]. In recent years, Bioinformatics, a branch of computational biology is emerging as a hottest research field among the Biotechnology and information technology researchers[3]. This field has risen in parallel with the developments of automated high *throughput* methods of biochemistry and biological discovery that yield a variety of forms of experimental data, such as DNA& RNA sequences, gene expressions patterns and chemical structures. The field's rapid growth is spurred by the vast potential for new understanding that can lead to new technological treatments, new agro-crops cultivation and new pharmaceutical drug discovery. In the recent years, most Bioengineering disciplines have started to adopt the information technology oriented curriculum due to its high performance computing, data interoperability, web-based platform compatibility and secured a suitable job opportunity[2,3]. This study discusses the functional research areas of life science that requires information technology platforms and highlights the scope, challenges and opportunities for computational biology in the Malaysian region.

Primary areas and functions: The following key areas of computational biology enumerate its major function:

Genomics:
Description: Genomics is one of the leading research areas in life science, which includes the range of research from the molecular basis of disease to biochemistry and to developmental biology.

Their main paradigms are: identification, cloning and analysis of a specific gene product for a given

Corresponding Author: R.J. A. Richard, Centre for Multimedia Computing, Faculty of Information Technology, Multimedia University, Cyberjaya, Malaysia, Ph: +603-8312 5407 Fax: +603-8312 5264

function is responsible for most of what we know in modern biology[6-8]. This creates an impact on identifying and cloning of new genes.

Scope and growth: High-throughput genomics technologies such as genome sequencing and whole-genome expression analysis is transforming the biological sciences now. The genome analysis and sequencing helps the scientists to identify similar RNA and DNA sequences and turn into cloning of new genes[6,7].

The extraordinary success of genomic technology, as well as the massive investment in genomics by governments and the pharmaceutical industry increased the job opportunities.

Bioinformatics:
Description: Bioinformatics is the new field of study that seeks to combine Biotechnology with Information Technology. Its basic tools (especially methods for finding relationships between genes) have already proved tremendously useful. Most of the Biological data banks, accessed tremendously by the researchers are dependent on these Bioinformatics tools[2,9].

Scope and Growth: It is a very young and incomplete field, whose deficiencies are a major anticlimax in the genomics revolution. Functional aspect of Bioinformatics is the representation, storage and distribution of data. Intelligent design of data formats and databases, creation of tools to query those databases and development of user interfaces that bring together different tools to allow the user to interact with the databanks through the web.

Proteomics:
Description: A relatively new area, proteomics studies, which do not rely on the entire genome, but rather on the portion of the genome that is, expressed in particular cells. This often involves cutting-edge technology, such as the usage of micro arrays, which allows the expression level of thousands of genes in a cell sample to be quickly determined[10].

Scope and growth: It targets the location of drug/gene therapy. Bioinformatics specialists work closely with bench scientists accomplish the "data mining" that lies behind this next wave of the pharmaceutical industry[10].

Sequence Gathering:
Description: The genome of an organism is assembled from thousands of fragments that must be correctly "stitched" together. The sequence analyzing tools and algorithms will help to merge them in sequences.

Scope and growth: Standardization of biological sequences into XML formats, simplified the process of sequence gathering and comparison by the cross references of biological data banks[2,11-12]. The growth

of BioXML, BSML, etc., assures the interoperability among the biological data banks.

Sequence Analysis:
Description: Sequence analysis is the most primitive operation in computational biology. This operation consists of finding which part of the biological sequences are alike and which part differs while medical analysis and genome mapping processes[13].

Scope and growth: Knowledge based single sequence analysis for sequence characteristics, pairwise sequence comparison and sequence-based searching, multiple sequence alignment and phylogenetic inference[10].

Database Maintenance:
Description: Many pharmaceutical companies and research institutes maintain private data banks of gene sequences and other biological and chemical information. These data banks have to achieve common formats.

Scope and growth: These repositories must be continually updated with data generated internally and from outside sources. This is a challenging task and the design and maintenance of these complex databases has become an important part of these complex databases have become an important part of Bioinformatics.

Pharmacogenomics:
Description: Pharmacogenomics is the application of genomic approaches and technologies for the identification of drug targets.

Scope and growth: New drug discovery using Bioinformatics tools and disease control.

There are other emerging fields using Bioinformatics tools such as Chemeoinformatics, Agro-informatics, Medical-informatics, etc.

Challenges in Computational Biology: The foremost challenge lies in computational biology is the understanding of the structure of a cell. A cell which contains hundreds of thousands of proteins that depends not only on the linear sequence of amino acids, presence of fats, sugar and water in its micro environment, but also on other molecules in their immediate proximity[10]. Complete understanding of the proteome will require new supercomputer architectures; immensely large databases, new data mining methods, new modeling and simulation techniques and the networking expertise to integrate data from desperate sources such in such a manner that it should be effective and affordable[14].

In this present trend, the rate of change in computer-enabled technological innovation is accelerating and hence practical applications of computing to unravel the proteome and other bioinformatics challenges are growing exponentially. In this regard, Bioinformatics should be given a high

priority by programmers, systems architects and other computer technology professionals to make an opportunity to take a proactive role in defining and shaping not only their future, but the future of humanity as well.

Challenges in the Sequence Analysis: The primary challenge for the sequential analysis of genomics is the existence of large volumes of data[10]. It must provide an analysis that can keep up and decipher the inherent structure of information within the data. This is due to the nature of the data that are detailed, complex and voluminous, representing the complete genetic blueprint for a living organism.

The Challenges of Genomics: Genomics is based on automated, high-throughput methods for generating experimental data, ranging from gene sequence, to gene expression levels, to protein-protein interactions[6-7,10]. Genomics technologies are making it possible to study biology as a system rather than as a microcosm. The completeness of the dataset and the systematic nature of the experiments, is designed to enable total analysis of the system and explicit comparison of any pair or any number of different experiments. These characteristics are the fundamental foundation for the next era in biology reassembling the understanding of molecular pieces into predictive models of the biology as a composite system.

Developing a quantitative analysis and understanding of the genome is very complex. At the very basic level, it is very difficult to identify the unknown genes through computer analysis of genomic sequences. The biological data and sequences are also very complex and interlinked with each other. A spot on a DNA Array, for instance, is connected not only to immediate information about its intensity, but two layers of information about genomic location, DNA sequence, structure, functions, etc. To find a metallic and biochemical pathway of genes representation in genomes will need collaborative work of computational biologists, computer scientists, mathematicians and statisticians[10].

This challenge will define Bioinformatics as a science. It is the analysis component for making sense of genomics data and putting it together to make predictive models of biological systems.

Fundamentally, this is the real opportunity of Bioinformatics; its an interesting theory driven by the benefit of one of the biggest, richest waves of systematic complete data ever produced. Theory suffers when data are too sparse or weak to falsify or "confirm" predictions, the data do not challenge new theoretical frameworks, or fundamental knowledge or theoretical tools are missing[15].

Career opportunities: Recently, demand for trained and skilled personnel has opened up a new career for bioinformaticians, especially in research &

development sectors. This demand will clearly become exacerbated with advances in genomics and post genomic research.

The career in Bioinformatics is divided into two parts such as: developing software and utilizing it. Bioinformatics graduates are in high demand. Career opportunities in Bioinformatics will continue to expand on both the computer and biological science fronts. In computer science, there are opportunities in programming, database development, systems analysis and software engineering. In the biological sciences career opportunities exist for Bioinformatics experts who assist biologists and medical researchers to interpret biological or medical data, design user interfaces and run laboratory information management and analysis systems. Opportunities also exist in research laboratories in universities, hospitals and the biotechnological and pharmaceutical industries[16,17].

Feasibility study: Curriculum on computational biology-A Malaysian perspective: Recently computational biology curriculum has been started in universities at postgraduate level to produce bioinformaticians and computational biologists. Malaysian government's initiatives in Bioinformatics will give more opportunities computational biologists, Bioinformatics researchers and bioinformaticians. Our study of computational biology explains the importance of this emerging technology. We would like to list down the findings of this study:

* Computational biology, though seems to be the hottest research area, has to reach the student community yet.
* It has rapid growth, which results in higher job opportunities.
* Merging of information technology to life sciences will bring more job opportunities. Information Technology concepts and subjects should be included in the Life science curriculum to fulfill the future requirements.
* Mathematics and Statistics knowledge are important to analyze the biological sequences. Both areas to be covered while developing computational biology curriculum.
* Governments and industries have initiated research and development programs in the life science area and this emerging field will be the most wanted technology in future.
* Currently most universities in Malaysia have a Life Sciences (Biology, Biotechnology, Molecular Biology, Agriculture, etc.) and Information Technology courses at undergraduate and postgraduate level. Thus they have the infrastructure to start computational biology oriented curriculum at the undergraduate level.
* With this we could say, Malaysia has consistent development in biotechnology and the Government of Malaysia has allotted funds for Biotechnology

field. Since Asia-Pacific region has produced rapid growth in this field, Malaysia can take a lead in Bioinformatics in Asia-Pacific region, by collaborating research organizations, academic sectors and industries[18]. According to its vision 2020, the country needs more trained computational biologist and Bioinformatics to fulfill the job requirements. The country is encouraging researchers to work on Agro-informatics, which is focusing on seed technology, new crop cultivation, tissue culture and other researches on the agricultural sector. The Malaysian Government has initiated Bio Valley project to work on Biotechnology and Bioinformatics.

BioValley Malaysia: Targets and goals: Proposed under the Eighth Malaysia Plan, BioValley Malaysia will be the nucleus of the country's biotechnology industry. It adopts a cluster approach that has been successfully applied to develop biotechnology industries around the world. The cluster will comprises biotechnology research institutions, universities and specialized companies. This is to foster inter-relationships and the convergence of intellectual expertise and entrepreneurship as well as the sharing of resources, which will enable the industry to establish a center of excellence. BioValley will integrate existing research resources with new facilities, equipment and human resources that will seed growth and innovation. BioValley will stimulate the growth and application of biotechnology in Malaysia by catalyzing research through the formation of new Institutes and linking Malaysian academic centers to industry.

NBBnet: National Biotechnology and Bioinformatics Network (NBBnet) is an initiative by the Malaysian government. It gives R&D support to the researchers for conducting Biotechnology and Bioinformatics researches. It allows access to biological data banks and other resources to conduct better researches in this hottest research area. Research institutes and the center of excellences can use these available resources[19].

Malaysia's initiatives in biotechnology will bring the nation to be a top leader in Biotechnology and Bioinformatics in Asia Pacific region. This emerging field will require trained Bioinformaticians and Computational Biologists to fulfill the requirements of the industry. Since computational biology has rapid growth.

In the Asia Pacific region, we strongly feel universities can take initiatives to start up the Bioinformatics curriculum at Under Graduate level to give a strong foundation in Genomics, Proteomics, Bioinformatics tools and Computational tools for Molecular Biology. Also universities can focus on certain areas according to their specialization, but common courses can be introduced to give basic awareness.

Proposed computational biology curriculum: Recent developments in computational and related fields have led to an explosive growth in biological information. Biological data are being generated faster than it can be analyzed and utilized. The importance of this field to drug discovery, biotechnology, agro-informatics, tissue culture, crop cultivation has resulted in a rapidly emerging commercial Bioinformatics sector. There is a growing trend for professionals with a strong foundation in both information technology and life sciences. The current demand is very high; we strongly feel curriculum on computational biology at under graduation level will educate students with expertise in both computing and life sciences and who have learned to integrate these two disciplines[20,21]. With the above discussion, we would recommend the courses below for the three years computational biology curriculum.

Year-1 Semester-1
Introduction to Molecular Biology
Basic Mathematics for Computational Biology
Introduction to Computer Programming Languages (C & BioPerl)
Database systems

Year-1 Semester-2
Bioinformatics Computing
Statistics and Probability theory

Year-1 Semester-3
Biophysical Chemistry (I)
Introduction to sequence analysis
Representations and Algorithms in Computational Biology

Year-2 Semester-1
Programming in Object Oriented Languages
Genetics, Genetic Information and Flow Processing
Protein and Nucleic Acid Structure, Dynamics and Engineering
Biophysical Chemistry (II)

Year-2 Semester-2
Computer Graphics and Visualization in Computational biology
Data Mining and machine learning techniques

Year-2 Semester-3
Introduction to system and computational Neuroscience
Biomaterials
Genetics, Evolution and Behavior
Study of Human Genome

Year-3 Semester-1
Industrial Training

Year-3 Semester-1
Elective- 1
Elective -2
FYP-1

Table1: Optional subjects

Course	Targeted Students
Basic Biology (Non Biology Students)	Non Biology students
Human-Computer Interaction	All
Molecular Biology of Proteins	All
Software Engineering	All
Agro informatics	All
Computer Communication Networks	All
Neural Networks	All
Bio Chemistry and Metabolic Pathways	All
Study of Human Genome	All
Structural Biology	Non Biology Students

Year-3 Semester-3
Artificial Intelligence
Advanced Computational Biology
Elective 3
FYP-2
(FYP- Final Year Project)

Optional Subjects: Optional Subjects are listed down in Table1.

CONCLUSION

Computational Biology is a field of science at the cutting edge of two of the fastest growing areas of technology information technology (computer science) and biotechnology (genetics, molecular biology and biochemistry). It is an interface between computing, mathematics, statistics and the Life Sciences and it is responsible for producing new computer systems (databases, software, networks and even hardware) and for using computer systems to solve problems in a wide variety of areas ranging from biology to medicine. The idea of developing/introducing the new curriculum on computational biology is:

* To provide the necessary awareness on computational biology.
* To provide infrastructure and resources for Bioinformatics and biotechnology information.
* To produce skilled computational biologists and bioinformaticians for the future requirements and to promote the development of a research program.
* To provide Bioinformatics tools and resources such as databases, software, network communication, training and consultancy.

REFERENCES

1. Mesirov, J.P. and D.K. Slonim, 1999. Computational Biology. Computing in Sci. and Engg., 1: 16-17.
2. Richard, R.J.A., N. Sriraam, A. Arokiasamy and A.M. Mat Jais, 2003. Perspectives of XML in Bioinformatics applications. Proc. Intl. Symp. Commune. Technol., Malaysia, pp: 80-83.
3. Liebman, M.N., 2001. Bioinformatics, Engineering in Medicine and Biology Society. Proc. 23rd Ann. Intl. Conf. IEEE, 4: 4063.
4. Doom, T., M. Raymer and D. Krane, 2004. Bioinformatics Potentials. IEEE, 23: 24-27.
5. Doom, T., M. Raymer, D. Krane and O. Garcia, 2003. Crossing the interdisciplinary barrier: A Baccalaureate Computer Science option in Bioinformatics. IEEE Trans. Education, 46: 387-393.
6. Moore, S.K., 2000. Understanding the human Genome, Spectrum. IEEE, 37: 33-35.
7. Genome Sequencing, <http://genome.gsc.riken.go.jp/hgmis/faq/seqfacts.html>
8. William, H.E. and J. Zobel, 2002. Indexing and retrieval of genomic database. IEEE Trans. Knowledge and Data Engg., 14: 63-78.
9. Bryan Bergeron, M.D., 2002. Bioinformatics Computing. Prentice Hall, pp: 45-47.
10. Cynthia Gibas and Per Jambeck, 2001. Developing Bioinformatics Computer Skills. O'Reilly and Associates, Inc., pp: 1-44, 159-160, 295-328.
11. Richard, R.J.A. and N. Srirram, 2004. XML Standards and Bioinformatics: An Overview. Asian J. Inform. Technol., 3: 11.
12. Sargunan, K.S., K.S.A. Kaur, R. Kumaresan, N. Sriraam, R.J.A. Richard and S. Nappis, 2004. Bioinformatics and data integration-An XML approach. Proc. Intl. Conf. Inform. and Commun. Technol.: From Theory to Applications, pp: 49-50.
13. Setubal, M., 2004. Introduction to Computational Molecular Biology. PWS Publishing Company, pp: 47.
14. Heath, L.S. and N. Ramakrishnan, 2002. The emerging landscape of Bioinformatics software systems. IEEE Computer, 35: 41-45.
15. See-Kiong Ng. and Limsoon Wong, 2004. Accomplishments and Challenges in Bioinformatics. IT Professional, 6: 44-50.
16. Bioinformatics opportunities, www.bioinformatics.org <http://www.bioinformatics.org>
17. Employment Opportunities in Bioinformatics, http://www.geocities.com/bioinformaticsweb/employement.html.
18. Miyano, S. and S. Ranganathan, 2001. The Asia-Pacific regional perspectives on Bioinformatics. IEEE Intelligent Systems, 16: 19-61.
19. Biovalley Malaysia: www.nbbnet.gov.my
20. Hughey, R. and K. Karplus, 2001. Bioinformatics: A new field in engineering education. Proc. Frontiers in Education Conf., 2 F2B: 15-19.
21. Doom, T., M. Raymer, D. Krane and O. Garcia, 2002. A proposed undergraduate Bioinformatics curriculum for computer scientists. Proc. 2002 ACM Special Interest Group on Computer Science Education (SIGCSE 2002), Convington (KY), 34: 78-84.